Bayesian instrumental variables: priors and likelihoods*

Hedibert F. Lopes and Nicholas G. Polson

University of Chicago

June 28, 2012

Abstract

Instrumental variable (IV) regression provides a number of statistical challenges due to the shape of the likelihood. We review the main Bayesian literature on instrumental variables and highlight these pathologies. We discuss Jeffreys priors, the connection to the errors-in-the-variables problems and more general error distributions. We propose, as an alternative to the inverted Wishart prior, a new Cholesky-based prior for the covariance matrix of the errors in IV regressions. We argue that this prior is more flexible and more robust thanthe inverted Wishart prior since it is not based on only one tightness parameter and therefore can be more informative about certain components of the covariance matrix and less informative about others. We show how prior-posterior inference can be formulated in a Gibbs sampler and compare its performance in the weak instruments case for synthetic as well as two illustrations based on well-known real data.

Keywords: Bayesian Learning, IV regression, Errors-in-Variables, Fat-tails, Inverted Wishart, Cholesky Decomposition, Demand for cigarettes, Angrist-Krueger data.

*Hedibert F. Lopes is Associate Professor of Econometrics and Statistics and Nicholas G. Polson is Professor of Econometrics and Statistics, Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave, Chicago, IL 60637, [hlopes,ngp]@chicagobooth.edu.

1 Introduction

Simultaneous equation models (SEMs) (Zellner, 1971, Chetty, 1966) and Instrumental Variables (IV) are fundamental tools in Statistics and Econometrics. Additional Bayesian work on SEM are Drèze (1976), Drèze and Morales (1976), Drèze and Richard (1983) and Kleibergen (1998), amongst others. IV regression has been tackled with a plethora of methods including: Bayesian approaches (Lindley and El-Sayyad, 1968, Zellner, 1971, Kleibergen and van Dijk, 2007), Bayes-Stein shrinkage (Zellner and Vandaele, 1975), decision-theoretic methods (Chamberlain, 2007), method of moments (Zellner, Tobias and Ryu, 2007), semi-parametric Dirichlet mixtures (Conley, Hansen, McCulloch and Rossi, 2008, Florens and Simoni, 2010) and Monte Carlo simulation (Zellner, Bauwens and van Dijk, 1988), to name but a few. Comparisons between Bayesian and classical approaches have been discussed in Lindley and El-Sayyad (1968) and Kleibergen and Zivot (2003). This problem is intertwined with that of "errors-in-the-variables" models (Zellner, 1971, Minka, 1999), co-integration (Stachan, 2003, Kleibergen and Paap, 2002, Villani, 2005, Koop, Leon-Gonzalez and Strachan, 2010) and reduced rank regression (Geweke, 1996). For a discussion of Bayesian approaches to IV and many examples in Economics and Marketing, see Lancaster (2004) and Rossi, Allenby and McCulloch (2005).

In this paper we will revisit and discuss key formulation, identification and estimation aspects when performing Bayesian inference in the instrumental variable regression model based on several of the above alternative representations. As opposed to simple linear regression models, Sims (2007), for instance, highlights the many issues with a priori assumptions in modeling the IV system as inferences can be very sensitivity to these specifications in the reduced form model. In addition, a common feature of many of these IV problem is that you can obtain unexpected "sharp" (and possibly ill-behaved) posterior distributions from "weak" prior distributions, mainly due to the heavy tails of the likelihood or the nonlinearity of the parameters of interest or both (Maddala, 1976, Zellner, 1971, Hoogerheide, Kaashoek and H. K. van Dijk, 2007, Hoogerheide and van Dijk, 2008a,b, Hoogerheide, Kleibergen and

van Dijk, 2008).

We propose a new Cholesky-based prior for the covariance matrix of the errors in IV regressions, as an alternative to the inverted Wishart prior. Rossi *et al.* (2005) highlight the importance of priors on the error covariance-matrix Σ as one goes from the structural model to the reduced form (see also Lancaster, 2004). Rossi *et al.* (2005, page 29) point out to several of the drawbacks of the Wishart distribution: "The most important is that the Wishart has only one tightness parameter. This means that we cannot be very informative on some elements of the covariance matrix and less informative on others." We argue that our Cholesky-based prior is more flexible and avoids such drawbacks.

The remainder of the paper is organized as follows. The basic IV regression set up and its Bayesian solution are outlined in Section 2. Important departures, such as more general, noninformative prior specifications and more general error distributions, are presented in Section 3, where we also describe the prior sensitivity issues raised in Sims (2007). Section 4 makes the connection between IV regressions and the likelihood-based approach of Zellner and Minka in the "errors-in-the-variables" models (EVM). Our Cholesky-based prior for the simultaneous error covariance is introduced and discussed in Section 5. The section includes three illustration: one based on simulated data and two based on well-known IV studies. Section 6 concludes.

2 Instrumental variables regression

The basic set-up is as follows which is drawn from Rossi *et al.* (2005). Let y_i be the response variable and x_i the (endogenous) regressor obeying, for $i = 1, \ldots, n$, the system of equations:

$$x_i = z_i'\delta + \varepsilon_{1i} \tag{1}$$

$$y_i = \gamma + \beta x_i + \varepsilon_{2i}, \tag{2}$$

with z_i a p-dimensional vector of instruments, related to x_i but independent of ε_{2i}. For simplicity an intercept is included in z, such that there are, in fact, only $p-1$ instrumental

variables in the above structure. We will assume that $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})'$ are i.i.d. $N(0, \Sigma)$, i.e. a bivariate normal distribution with zero mean vector and Σ variance-covariance matrix, where Σ has diagonal components σ_{11} and σ_{22} and off-diagonal component $\sigma_{12} = \rho(\sigma_{11}\sigma_{22})^{1/2}$. More general error structures are discussed in Section 3.1.

The key distinction between the above system of equations to a standard bivariate regression is the possible correlation between the error terms ε_{1i} and ε_{2i} (and, therefore, between x_i and ε_{2i}). This leads to the well-known "endogeneity" bias when learning β from Equation 2; that is the information of x_i that is correlated with ε_{2i} should not be used when learning about the regression parameter β.

The reduced form representation of Equations 1 and 2 is

$$x_i = z_i'\pi_x + \nu_{1i} \tag{3}$$

$$y_i = \gamma + z_i'\pi_y + \nu_{2i}, \tag{4}$$

where $\pi_x = \delta$, $\nu_{1i} = \varepsilon_{1i}$, $\pi_y = \beta\delta$ and $\nu_{2i} = \beta\varepsilon_{1i} + \varepsilon_{2i}$. The relation between ε_i and ν_i is

$$\nu_i = \begin{pmatrix} 1 & 0 \\ \beta & 1 \end{pmatrix} \varepsilon_i = B\varepsilon_i. \tag{5}$$

The model is not identified in the limiting case of $\delta = 0$. More generally, if the instruments z explain only a small portion of the variability of x (weak instrument case), then the likelihood function is concentrated around $\beta + \sigma_{12}/\sigma_{11} = c$ for some estimable constant c.

Finally, it is worth noting that the interpretation of ρ as a measure of endogeneity needs to be extended when inference is performed from a Bayesian viewpoint. More specifically, the unconditional distribution of ε_i, once Σ is integrated out, might still exhibit dependence even when $\rho = 0$. Unconditionally, i.e. integrating out Σ, the dependence between ε_{1i} and ε_{2i} is actually higher than the conditional bivariate normal with a small ρ and potentially much higher when degrees of freedom for the prior is small. In this paper we will focus on what can be called "weak endogeneity in the frequentist sense", or conditional endogeneity, to make it explicit we are not considering the case where Σ is integrated out.

2.1 Posterior inference

Recall that ε_is are i.i.d. $N(0, \Sigma)$, such that the distribution of the reduced form errors ν_i is also bivariate normal $N(0, \Omega)$ where

$$\Omega = B\Sigma B' = \begin{pmatrix} \sigma_{11} & \beta\sigma_{11} + \sigma_{12} \\ \beta\sigma_{11} + \sigma_{12} & \beta^2\sigma_{11} + 2\beta\sigma_{12} + \sigma_{22} \end{pmatrix} \tag{6}$$

so that β and Σ are intertwined in the reduced form and independent priors for both parameters would be counter-intuitive.

We will start, instead, with the prior specification for the structural parameters from Equations 1 and 2 discussed in Rossi *et al.* (2005):

$$\delta \sim N(d_0, D_0) \tag{7}$$

$$(\gamma, \beta)' \sim N(b_0, B_0) \tag{8}$$

$$\Sigma \sim IW(v_0, \Sigma_0), \tag{9}$$

for known hyperparameters d_0, D_0, b_0, B_0, v_0 and Σ_0. $IW(v_0, \Sigma_0)$ standards for the Inverted-Wishart distribution with parameters v_0 (prior degrees of freedom) and Σ_0 (prior scale matrix), whose density is $p(\Sigma) \propto |\Sigma|^{-(v_0+p+1)/2} \exp\{-\frac{1}{2}\text{tr}\Sigma_0\Sigma^{-1}\}$ and $v_0 > p$. If $v_0 \geq p+2$, then $E(\Sigma) = \Sigma_0/(v_0 - p - 1)$. In this paper $p = 2$.

Sampling Σ. The full conditional distributions under the above specification are straightforward. First, the error variance has posterior

$$(\Sigma|\gamma, \beta, \delta, \text{data}) \sim IW(v_0 + n, \Sigma_0 + S),$$

where $S = \sum_{i=1}^{n} \varepsilon_i\varepsilon_i'$ and $\text{data} = \{(x_i, y_i, z_i); i = 1, \ldots, n\}$.

Sampling (γ, β). Secondly, the regression parameters (γ, β) have a joint distribution of the form

$$(\gamma, \beta|\delta, \Sigma, \text{data}) \sim N(b_1, B_1),$$

where

$$B_1^{-1} = B_0^{-1} + \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i' \quad \text{and} \quad B_1^{-1} b_1 = B_0^{-1} b_0 + \sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i,$$

for

$$\begin{aligned} \tilde{x}_i &= (1, x_i)'/\sigma_{2|1}^{1/2} \\ \tilde{y}_i &= (y_i - (x_i - z_i'\delta)\sigma_{12}/\sigma_{11})/\sigma_{2|1}^{1/2} \\ \sigma_{2|1} &= \sigma_{22}(1-\rho^2). \end{aligned}$$

Sampling δ. Finally, the regression parametes for the instrument, δ, have a distribution of the form

$$(\delta|\gamma, \beta, \Sigma, \text{data}) \sim N(d_1, D_1),$$

where

$$D_1^{-1} = D_0^{-1} + \sum_{i=1}^{n} \tilde{z}_i \tilde{z}_i' \quad \text{and} \quad D_1^{-1} d_1 = D_0^{-1} d_0 + \sum_{i=1}^{n} \tilde{z}_i \tilde{x}_i,$$

for

$$\begin{aligned} \tilde{x}_i &= (x_i - (y_i - \gamma - \beta x_i)\sigma_{12}/\sigma_{22})/\sigma_{1|2}^{1/2} \\ \tilde{z}_i &= z_i/\sigma_{1|2}^{1/2} \\ \sigma_{1|2} &= \sigma_{11}(1-\rho^2). \end{aligned}$$

2.2 Illustration

The performance of the above Gibbs sampler is illustrated with a data set with $n = 200$ observations simulated from Equations 1 and 2 with $\gamma = 1.0$, $\beta = 0.5$, $\sigma_{11} = \sigma_{22} = 1$, $\delta = (1.0, 0.1)'$ (one weak instrument) and $\rho = 0.1$ (low degree of endogeneity).

The prior hyperparameters are $d_0 = b_0 = 0$ and $D_0 = B_0 = 25I_2$, suggesting relatively low prior information about δ, β and γ. For Σ two priors are entertained: $v_0 = 3$ and $\Sigma_0 = 3I_2$ (relatively vague prior) and $v_0 = 0.00001$ and $\Sigma_0 = 0.00001I_2$ (non-informative prior). The distinction between "relatively vague" and "non-informative" is arbitrary and

its only purpose is to distinguish two prior specification for Σ. As it can be seen, the "non-informative" prior tries to mimic the improper Jeffreys prior for covariance matrices, while the "relatively vague" prior is proper and has prior mean but no prior variance.

Figure 1 summarizes our findings and shows that the variability of σ_{12}/σ_{11} and of β are greatly affected by the choice of the prior on Σ. Posterior inference based on the relatively vague prior (top row) turns out to be rather informative when compared to posterior inference based on the non-informative prior (bottom row). Notice that β and δ (or σ_{12}/σ_{11}) become relatively unrelated (top middle and right panels), suggesting again that the relatively vague prior turns out to be too informative regarding these linear dependences. Meanwhile, under the non-informative prior, β (similar for σ_{12}/σ_{11}) becomes more diffuse *a posteriori* (varies between $(-6, 6)$) when δ approaches zero (or the instrument becomes innocuous) highlighting the sharp behavior of the likelihood function in the vicinity of non-identifiability (bottom middle and right panels) . The pattern repeats itself, as expected, regardless of the sample size, since the (non-identified) likelihood dominates the non-informative prior even for fairly small sample sizes. These plots resemble the bivariate distributions with the conditional normal example of Arnold and Strauss (1991) and the example of uncorrelated models with normal marginals of Ebrahimi *et al.* (2010). Comparisons of the top and bottom panels show that the parameters of Wishart prior can have profound distributional effects. See also the discussion on Section 5 of Hoogerheide and van Dijk (2008a) where similar plots for (β, δ) are exhibit.

To sum up, the trade-off between more precise (and less accurate) and less precise (and more accurate) posterior inference is an important, problem-specific part of the modeling process that needs to be dealt with on a case-by-case basis. In the next section we discuss alternatives to the above model and prior specifications.

3 More on IV regression

A number of authors have proposed the use of Jeffreys prior in the IV regression. From the reduced form (Equations 3 and 4), let $\pi = (\pi_x, \pi_y) = (\delta, \beta\delta)$ a $(p \times 2)$ matrix of rank one. The Jeffreys prior is then given by

$$\left| \frac{\partial \pi}{\partial(\beta, \delta)} \left(\frac{\partial \pi}{\partial(\beta, \delta)} \right)' \right| = ||\delta|| (1 + \beta^2)^{\frac{1}{2}}, \tag{10}$$

which is suggestive of using priors with polynomial tails, such as a Cauchy (Chao and Phillips, 1998, and Sims, 2007).

The reduced form IV model where inference for $\pi_y = \beta\delta$ is required (see text after Equation 4) is related to the analysis of a product of two normal means (Lindley and El-Sayyad, 1968, Berger and Bernardo, 1989). The issue is two-fold, we have a high dimensional parameter vector that requires prior regularisation and we have to learn how much shrinkage to employ. There is a fundamental trade-off in the fit of the two regression equation described above. This was illustrated by the example in the previous section and presented in Figure 1.

Sims (2007) makes a number of important remarks about prior sensitivity in IV problems. Two of them are as follows.

- *As the likelihood does not go to zero as $\beta \to \infty$, with Σ fixed, no matter what the sample size is there is an issue of prior sensitivity.*

  Assume we combine a likelihood with a prior that has Gaussian tails. As the likelihood mode (maximum likelihood estimator) moves away from the prior mean (and mode), the posterior mean (and mode) will move away from the prior mean (and mode) at first (as expected), but then reverses the direction, coming back to settle at the prior mean (or mode) even when the likelihood mode and prior mean (or mode) are well separated. These effects of fat-tailed combinations of likelihoods and priors has been documented in the linear Bayes regression model by West (1984) and exploited to find sparse estimators by Carvalho, Polson and Scott (2010).

- *Put another way, in a sample where the posterior is highly non-Gaussian and has substantial, slowly declining tails, even apparently weak prior information* $(\delta, \beta) \sim N(0, 100I)$ *can substantially affect the inference.*

  One approach then is to make the prior flexible enough to accommodate fat-tails and to *learn* from the data how thick these tails should be, see Lopes and Polson (2011). There will be an interaction with the specification of the prior magnitude on Σ. See the discussion in Lindley and El-Sayyad (1968) in their Bayesian treatment. Our approach will be to free up the prior on Σ and use a Cholesky-based prior rather than the standard inverse Wishart.

3.1 Dirichlet process prior

Conley *et al.* (2008) develop a Bayesian semi-parametric approach to the IV regression problem. They use a normal-based Dirichlet process prior to jointly model structural and instrumental variable equations errors (the errors in our Equations 1 and 2). More specifically, $\varepsilon_i \sim N(0, \Sigma_i)$ replaces the standard $\varepsilon_i \sim N(0, \Sigma)$, with Σ_i now i.i.d. from the discrete random distribution G, which in turn is modeled by a Dirichlet process with concentration parameter α and base distribution G_0, commonly denoted by $G \sim DP(\alpha, G_0)$. The marginal distribution of Σ_i (integrating out G) is continuous and is called a mixture of Dirichlet Processes (MDP). See Escobar and West (1995, 1998) for Bayesian posterior inference via MCMC for this class of models.

Conley *et al.* performed extensive sampling experiments and showed that, when the errors are actually normally distributed, their prior specification leads to more efficient results when compared to standard Bayesian (Section 2) and classical methods. They say, in their conclusion, that "Our Bayesian semi-parametric procedure produces credibility regions which are dramatically shorter than confidence intervals based on the weak instrument asymptotics. The shorter intervals from our method are produced by more efficient use of sample information." They go on and finish the paper saying that "$\cdots$ our non-parametric Bayesian method

dominates Bayesian methods based on normal errors and may be preferable to methods from the recent weak instruments literature if the investigator is willing to trade-off lower coverage for dramatically smaller intervals." To sum up, their Bayesian methodology should be the preferred one in many IV problems from here on.

3.2 Bayesian model averaging

Recent research has focussed on applying Bayesian model averaging across sets of instruments, exogeneity restrictions, the validity of identifying restrictions and the set of exogenous regressors are explored by Eicher, Lenkoski and Raftery (2009) and Koop, Leon-Gonzalez and Strachan (2011). Another avenue, which we do not explore here, is to assume a flexible fat-tailed alternative such as a mixture of t_ν distributions (mixed over ν). Our use of Cholesky priors for Σ follows through in these settings as well. As we will see there can be interesting sensitivity issues to the prior specification.

4 Errors-in-the-variables models

Minka (1999) proposed a proper Bayes approach to linear regression with errors in both equations. This builds on the well-known work of Zellner (1971, chapter 5). Prior spefication is an important feature of this problem, as Minka observes *Statisticians have worked on regression but they often shoot themselves in the foot by using priors that are too weak or too strong*. Our flexible cholesky-based prior on Σ will allieviate the problem.

The issues can be seen in the simple one dimensional case. The EVM is given by

$$x_i = z_i + \varepsilon_{1i} \tag{11}$$

$$y_i = \beta z_i + \varepsilon_{2i}, \tag{12}$$

where we take $\delta = 1$ and $\gamma = 0$ from the previous specification. The $z_1, \ldots, z_n$s are now latent unobserved variables, or Zellner's "incidental parameters", which we endow with a

normal prior $z_i \sim N(0, \tau^2)$. The covariance of the error term is again Σ, which, for the purpose of illustration, will be considered diagonal, that is $\Sigma = \text{diag}(\sigma_{11}^2, \sigma_{22}^2)$.

Let $(x, y) = (x_1, \ldots, x_n, y_1, \ldots, y_n)$ be the observed data and $S = \sum_{i=1}^{n}(x_i, y_i)'(x_i, y_i)/n$ be the sample covariance matrix. Zellner (1971) was the first to calculate the Bayes marginal likelihood given by

$$p(x, y|\theta, \Sigma, \tau^2) \propto |\Sigma|^{-\frac{n}{2}} \left(1 + \tau^2 \theta' \Sigma^{-1} \theta\right)^{-\frac{1}{2}} \exp\{-0.5\text{tr}(GS)\}, \tag{13}$$

where $\theta = (1, \beta)'$ and $G = \Sigma^{-1} - \Sigma^{-1}\theta \left(\theta'\Sigma^{-1}\theta + \tau^{-2}\right)^{-1} \theta'\Sigma^{-1}$.

Zellner's Bayes marginal likelihood approach can differ dramatically when compared to the conditional two stage least squares (2SLS) approach. Here, if q is the principal eigenvector of $S\Sigma^{-1}$ then the MLE is $\widehat{\beta} = q_2/q_1$ and the eigenvalues play a minor role conditionally. In the marginal Bayes likelihood, however, they have a very influential effect and it is typical for the marginal likelihood to have two finite local maxima. The case where they agree is when $\tau \to \infty$ although this is rarely the case in practice. Another special case of interest is when the largest eigenvalue is small. Then the marginal likelihood has one maximum and it is near zero and all the variation in the data can be explained by noise and $\widehat{\beta}$ is driven towards zero.

5 Cholesky-based prior

This section will fouces on one choice equation and one outcome equation although it can directly be extended to multivariate scenarios. The use of Cholesky-based priors is not new and they have had success in many situations; see, for example, Lopes, McCulloch and Tsay (2011) for an application to high dimensional stochatic volatility modeling (see also Pourahmadi, 1999, for longitudinal models).

The key idea is the following. Instead of modeling Σ via an inverted Wishart distribution with parameters v_0 and Σ_0, i.e. $\Sigma \sim IW(v_0, \Sigma_0)$, we will model the components of the

recursive conditional regressions that arises form the Cholesky decomposition of Σ. More precisely, recall from Section 2.1 that $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})'$ are i.i.d. $N(0, \Sigma)$ and let

$$\Sigma = AHA' \tag{14}$$

be the Cholesky decomposition of Σ such that A is lower triangular with ones in the main diagonal and lower triangular component given by $a_{21} = \sigma_{12}/\sigma_{11}$ and $H = \text{diag}(\sigma_{11}, \sigma_{2|1})$. The reverse transformation is given by $\sigma_{12} = a_{21}\sigma_{11}$ and $\sigma_{22} = \sigma_{2|1} + \sigma_{12}^2/\sigma_{11}$. Therefore

$$A^{-1}\varepsilon_i \sim N(0, H), \tag{15}$$

and $\varepsilon_i \sim N(0, \Sigma)$ can be rewritten by the following recursive conditional regressions (or simply *triangular regressions*)

$$\varepsilon_{1i} \sim N(0, \sigma_{11}) \tag{16}$$

$$\varepsilon_{2i}|\varepsilon_{1i} \sim N(a_{21}\varepsilon_{1i}, \sigma_{2|1}). \tag{17}$$

The parameter a_{21} measures the strength of the correlation between ε_{1i} and ε_{2i}, while $\sigma_{2|1}$ is the conditional residual variance. We then specify independent prior distributions for σ_{11}, a_{21} and $\sigma_{2|1}$. The implied prior for Σ can be directly obtained, either analytically or via Monte Carlo simulation. More specifically, σ_{11} is learned from equation (16), σ_{12} is learned from σ_{11} and a_{21} $(= \sigma_{12}/\sigma_{11})$, see equation (17), and σ_{22} is learned from σ_{11}, σ_{12} and $\sigma_{2|1}$ $(= \sigma_{22} - \sigma_{12}^2/\sigma_{11})$, also from equation (17).

The main attractiveness of the Cholesky-based prior is its relative freedom to independently quantity the uncertainty for the individual components of Σ, which turns out to be one of the major constraints of the Wishart and inverted Wishart distributions. See Rossi *et al.* (2005) and Lopes *et al.* (2011) for additional discussion on the limitations of the Wishart distribution.

The MCMC scheme for the IV regression model with Cholesky-based prior is the same as with the inverted-Wishart prior for all parameters but the covariance ones, that is σ_{11} and $\sigma_{2|1}$ and a_{21} (see Section 2.1). For σ_{11} and $\sigma_{2|1}$ we assign independent standard inverted

gamma priors, while a normal prior is specified for a_{21}. These priors are combined with equations (16) and (17) and, conditional on γ, β and δ, leads to standard Gibbs updates for σ_{11} and $\sigma_{2|1}$ and a_{21}. The next Section presents three illustrative applications.

5.1 Illustration 1 - synthetic data

We continue in the context of weak instrument and low endogeneity introduced in Section 2.2, that is $\delta = (1.0, 0.1)'$ and $\rho = 0.1$. Figure 2 compares the inverted Wishart prior, $\Sigma \sim IW(3, 3I_2)$, with the implied prior obained from the Cholesky-based prior of $(\sigma_{11}, a_{21}, \sigma_{2|1})$, where $\sigma_{11} \sim IG(0.75, 0.75)$, $\sigma_{2|1} \sim IG(3, 3)$ and $a_{21} \sim N(0, 0.7)$. The hyperparameters were selected to make both prior distributions on Σ as similar as possible. In addition, two prior specifications for (β, δ) are entertained. In the first case, $(\beta, \delta) \sim N(0, 25I_3)$, which represents diffuse but proper prior distributions. In the second case, $p(\beta, \delta) \propto ||\delta||(1 + \beta^2)^{\frac{1}{2}}$ (see equation 10), which can be thought of as a Jeffreys-type prior. This leads to four prior specifications for (β, δ, Σ). As before, the prior for γ is $N(0, 25)$.

The marginal posterior distributions for σ_{11}, σ_{22}, ρ and the joint marginal for $(\sigma_{12}/\sigma_{11}, \sigma_{22})$ and (β, δ) appear in Figures 3 and 4, respectively. The learning of σ_{11} and σ_{22} is similar across prior specifications, with the forth prior specification (Jeffreys for (β, δ) and Cholesky-based for Σ) leading to more informative posterior for σ_{22} as well as ρ and β.

Similar results are found when the priors on Σ or on $(\sigma_{11}, \sigma_{2|1}, a_{21})$ become extremely non-informative, that is, when $\Sigma \sim IW(0.00001, 0.00001I_2)$ or $\sigma_{11} \sim IG(0.000001, 0.000001)$, $\sigma_{2|1} \sim IG(0.000001, 0.000001)$ and $a_{21} \sim N(0, 1000)$. See Figures 5 and 6. It appears that the Jeffreys prior for (β, δ) when combined with a diffuse Cholesky-based prior for Σ leads to unstable results, while the other three combination produce relatively similar results.

5.2 Illustration 2 - Demand for cigarettes

Here we revisited the demand for cigarettes illustration presented in Chapter 10 of Stock and Watson's textbook *Introduction to Econometrics*, pages 339–341. The goal is to study the effect of price changes on the demand for cigarettes when using sales tax as an instrumental variable. The data set consists of annual data for the 48 continental U.S. states for the year of 1995 and the proxy for price is the logarithm of the average real price per pack of cigarettes including all taxes (x in our notation). In addition, the proxy for consumption is is the logarithm of the number of packs of cigarettes sold per capita in the state (y in our notation), while the proxy for sales tax is is the portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack in real dollars, deflated by the consumer price index (z in our notation). The previous description was taken from the webpage for Stock and Watson's book at `http://wps.aw.com/aw_stock_ie_2/50/13016/3332253.cw/index.html`.

Figure 7 presents the data with ordinary least squares estimates given by $(\gamma_{ols}, \beta_{ols}) = (10.850, -1.213)$ and $\delta_{ols} = (4.79006, 0.00667)$. The sample coefficient of correlation between the residuals of both OLS fits is around -0.1775306. These preliminary results suggest a scenario of a low degree of endogeneity and a relatively weak instrument, somewhat similar to the previous simulation exercise.

As with the simulation exercise, a fairly vague prior specification is used for all the model parameters, that is, $\Sigma \sim IW(a_0, a_0 I_2)$ or $\sigma_{11} \sim IG(a_0, a_0)$, $\sigma_{2|1} \sim IG(a_0, a_0)$, where $a_0 = 0.0000001$, and $a_{21} \sim N(0, \sigma_0^2)$, $(\beta, \delta) \sim N(0, \sigma_0^2 I_3)$, and γ is $N(0, \sigma_0^2)$, where $\sigma_0^2 = 1000000$. The initial values for the parameters are $\sigma_{11}^{(0)} = \sigma_{2|1}^{(0)} = 1$, $a_{21}^{(0)} = 0$, $\gamma^{(0)} = \beta^{(0)} = 0$ and $\delta^{(0)} = (0, 0)$. The performance of both MCMC schemes are presented in Figure 8. Both algorithms converge after several thousand draws, with the Cholesy-based one performing slightly better than the inverted Wishart one.

Posterior inference is summarized in Figure 9 and Table 1. The posterior median for the parameter that measure the degree of endogeneity, ρ, is -0.1934 while the posterior probability that $\rho > 0$ is about 10% (under both prior specifications). On the other hand, the instru-

mental variable, sales tax, has a nonzero effect of price since the 95% credibility interval for δ_1, $(0.0061, 0.0072)$, is away from zero. Based on the posterior of β, one can argue that an increase in the price of 1% reduces consumption by 0.67% to 1.59%.

5.3 Illustration 3 - Return to education

Here we revisited the return to education illustration presented in Chapter 8 of Lancaster's textbook *An Introduction to Modern Bayesian Econometrics*, pages 325-334. The goal is to study the effect of (years of) education on (log) wages when using quarter of birth as an instrumental variable. This is a fairly well known study and was first proposed by Angrist and Krueger (1991). We follow Lancaster's simplification and focus only on men born in 1939, which is the last year of the 10-year study of Angrist and Krueger (1991). They argued that quarter of birth might be related to years of education due to age-related regulations to both enter and leave school. More precisely, children whose birthdays fall in the 4th quarter of the calendar year will enter (elementary) school the fall of that same year or within one or two months from their birthdays, while children whose birthdays fall, say, in the 1st quarter of the calendar year will enter school at least six months after their birthday. In addition, compulsory schooling laws require students to remain in school until a predetermined age (usually sixteen or seventeen). Figure 10 shows the summary of both relationships. As it can be seen by the differences between the average years of education per quarter of birth, the instrument (quarter of birth) is relatively weak.

We perform Bayesian inference based on our Cholesky-based prior for the components of Σ. The prior hyperparameters specified as in the previous illustration, which leads to fairly vague prior information. Posterior summaries are presented in Figure 11. As expected, apart form δ_3, all δs are quite similar, corroborating with the initial suggestion that quarter of birth is a weak instrument for this illustration. Nonetheless, the posterior mean and median of the percentage marginal return, 100β, are 11.2% and 11.9%, respectively, while the 95% credibility interval is $(6.57\%, 16.5\%)$. Finally, the degree of endogeneity can be

measure by ρ, whose posterior mean and median are -0.245 and -0.191, respectively, while $Pr(\rho < 0|\text{data}) = 73.4\%$. Lancaster (page 333) says that "the posterior suggests that the structural form errors are negatively correlated and this is a bit surprising on the hypothesis that a major element of both ε_1 and ε_2 is *ability* and this variable tends to affect positively both education and wages. But the evidence is very far from conclusive." We agree and claim that this example, as well as the previous ones, illustrates the inferential difficulties when combining weak instruments and low degree of endogeneity.

6 Discussion

Instrumental variable likelihoods and their errors-in-the-variables cousin have challenging likelihood surfaces. With that comes the issue of sensitivity to prior specification. It has long been known that the Bayesian solution to the identifiability problem is attractive (Zellner, 1971, Conley *et al.*, 2008) and that, given a prior, inference can be based on marginal likelihoods.

Here we introduce the use of Cholesky-based priors, which are more flexible that the traditional normal inverse-Wishart regression priors, and more realistic than an "uninformative" Jeffreys prior. Recent work in standard regression problems has been addressed with heavy-tailed Cauchy priors (Gelman *et al.*, 2008). We show how prior-posterior inference can be formulated in a Gibbs sampler and compare its performance in the weak instruments case. Given modern-day computational methods for Bayesian inference (Gamerman and Lopes, 2006, Lopes, Carvalho, Johannes and Polson, 2011), these complicated likelihoods seem ripe for more discussion of prior sensitivity.

Acknowledgments

The authors would like to thank The University of Chicago Booth School of Business for providing financial support for our research. The authors are grateful to the Gues Editor Ehsan Soofi and the two anonymous referees whose invaluable comments and suggestions significantly improved the presentation and the quality of the paper. Finally, we would like to thank our beloved friend Arnold Zellner for being invariably enthusiastic about Bayesian statistics and an important source of inspiration and support.

7 References

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings. *Quartely Journal of Economics*, 106, 979-1014.

Arnold, B. C. and D. Strauss (1991). Bivariate distributions with conditionals in prescribed exponential families. *Journal of the Royal Statistical Society, Series B*, 53, 365-375.

Berger, J. O. and J. M. Bernardo (1989). Estimating the product of means: Bayesian analysis with reference priors. *Journal of American Statistical Association*, 84, 200-207.

Carvalho, C. M., N. G. Polson and J. G. Scott (2010). The Horseshoe estimator of sparse signals. *Biometrika*, 97(2), 465-480.

Chamberlain, G. (2007). Decision theory applied to an instrumental variables model. *Econometrica*, 75(3), 609-652.

Chao, J. C. and P. C. B. Phillips (1998). Posterior distribution in limited information analysis of the simultaneous equations model using Jeffreys prior. *Journal of Econometrics*, 87, 49-86.

Chetty, V. K. (1966). Bayesian analysis of some simultaneous equation models and specification errors. *Unpublished PhD Thesis*, University of Wisconsin, Madison.

Conley, T., C. Hansen, R. E. McCulloch and P. E. Rossi (2008). A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144, 276-305.

Drèze, J. H. and J. A. Morales (1976). Bayesian full information analysis of simultaneous equations. *Journal of American Statistical Association*, 71, 919-923.

Drèze, J. H. (1976). Bayesian limited information analysis of the simultaneous equations model. *Econometrica*, 44, 1045-1075.

Dréze, J. H. and Richard, J. F. (1983). Bayesian Analysis of Simultaneous Equations Systems. In *Handbook of Econometrics*, volume 1, edited by Z. Griliches and M.D. Intrilligator. Amsterdam: Elsevier Science.

Ebrahimi, N., G. G. Hamedani, E. S. Soofi and H. Volkmer (2010). A class of models for uncorrelated random variables. *Journal of Multivariate Analysis*, 101, 1859-1871.

Eicher, T. S., A. Lenkoski and A. E. Raftery (2009). Bayesian model averaging and endogeneity under model uncertainty: an application to development determinants. *Technical report*, University of Washington.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.

Escobar, M. D. and M. West (1998). Computing non-parametric hierarchical models. In: Dey, D., P. Müller and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, New York, pp. 1-22.

Florens, J.-P. and A . Simoni (2010). Nonparametric estimation of an instrumental regression: a quasi-Bayesian approach based on regularized posterior. *Technical report*, Toulouse School of Economics.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall/CRC, Baton-Rouge.

Gelman, A., A. Jakulin, M. G. Pittau and Y.-S. Su (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2, 1360-1383.

Geweke, J. (1996). Bayesian reduced rank regression. *Journal of Econometrics*, 75, 121-146.

Hasegawa, H. and H. Kozumi (2001). Bayesian Analysis on Engel Curves estimation with measurement errors and an instrumental variable. *Journal of Business and Economic Statistics*, 19, 292-298.

Hoogerheide, L. F., J. F. Kaashoek and H. K. van Dijk (2007). On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank. *Journal of Econometrics*, 139, 154-180.

Hoogerheide, L. F., F. Kleibergen and and H. K. van Dijk (2008). Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. *Journal of Econometrics*, 138, 63-103.

Hoogerheide, L. F. and H. K. van Dijk (2008a). Possibly Ill-behaved Posteriors in Econometric Models. *Technical report*, Tinbergen Institute.

Hoogerheide, L. F., and H. K. van Dijk (2008b). Simulation-based Bayesian econometric inference. *Technical report*, Tinbergen Institute.

Kleibergen, F. and R. Paap (2002). Priors, posteriors, and Bayes factors for a Bayesian analysis of co-integration. *Journal of Econometrics*, 111, 223-249.

Kleibergen, F. and H. K. van Dijk (1994). On the shape of the likelihood and posterior in co-integration models. *Econometric Theory*, 10, 514-551.

Kleibergen, F. and H. K. van Dijk (1998). Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14, 701-743.

Kleibergen, F. and H. K. van Dijk (2007). Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. *Journal of Econometrics*, 138, 63-103.

Kleibergen, F. and E. Zivot (2003). Bayesian and Classical approaches to Instrumental Variable regression. *Journal of Econometrics*, 114, 29-72.

Koop, G., R. Leon-Gonzalez and R. Strachan (2010). Efficient posterior simulation for cointegrated models with priors on the cointegration space. *Econometric Reviews*, 29, 224-242.

Koop, G., R. Leon-Gonzalez and R. Strachan (2011). Bayesian Model averaging in the Instrumental variable regression model. *Technical report*, University of Strathclyde.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Blackwell Publishing.

Lindley, D. V. and G. M. El Sayyad (1968). The Bayesian estimation of a linear functional relationship. *Journal of Royal Statistical Society, Series B*, 30, 190-202.

Lopes, H. F., C. M. Carvalho, N. G. Polson and M. Johannes (2011). Particle learning for sequential Bayesian computation (with discussion). In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M and West, M., editors, *Bayesian Statistics 9*, 317-360.

Lopes, H. F. and N. G. Polson (2011). Particle learning for fat-tailed distributions. *Technical report*, The University of Chicago Booth School of Business.

Lopes, H. F., R. E. McCulloch and R. E. Tsay (2011). Cholesky stochastic volatility. *Technical report*, The University of Chicago Booth School of Business.

Maddala, G. S. (1976). Weak priors and sharp posteriors in simultaneous equation models. *Econometrica*, 44, 345-351.

Minka, T. (1999). Linear regression with errors in both variables: a proper Bayesian approach. *MIT Media Lab Note (10/8/99)*.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86, 677-690.

Rossi, P. E., G. M. Allenby and R. E. McCulloch (2008). *Bayesian Statistics and Marketing*. New York: Wiley.

Sims, C. (2007). Thinking about instrumental variables. *Technical report*, Princeton University.

Startz, R., C. R. Nelson and E. Zivot (1999). Improved Inference for the Instrumental Variable Estimator. In: *Econometric theory and practice*, 125-163.

Stock, J. H. and Watson, M. W. (2003). *Introduction to Econometrics*. Boston: Addison Wesley.

Strachan, R. W. (2003). Valid Bayesian estimation of the cointegrating error correction model. *Journal of Business & Economic Statistics*, 21, 185-195.

Villani, M. (2010). Bayesian reference analysis of cointegration. *Econometric Theory*, 21, 326-357.

West, M. (1984). Oultier models and prior distributions in Bayesian linear regression. *Journal of Royal Statistical Society, Series B*, 46, 431-439.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Zellner, A. and W. Vandaele (1975). Bayes-Stein estimators for k-means, regression and simultaneous equation models. In: S. E. Fienberg and A. Zellner, Editors, *Studies in Bayesian econometrics and statistics*, North-Holland, Amsterdam (1975), pp. 627-653.

Zellner, A. (1979). Statistical analysis of Econometric models. *Journal of the American Statistical Association*, 74, 628-651.

Zellner, A., L. Bauwens and H. K. van Dijk (1988). Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods. *Journal of Econometrics*, 38, 39-72.

Zellner, A., J. Tobias and H. K. Ryu (1997). Bayesian method of moments (BMOM) analysis of parametric and semi-parametric regression models. *Technical report*, University of Chicago.

Parameter	Median	95% credible interval
σ_{11}	0.0012	(0.0008,0.0019)
σ_{22}	0.0374	(0.0254,0.0584)
σ_{12}	-0.0013	(-0.0038,0.0008)
ρ	-0.1934	(-0.4655,0.1111)
δ_0	4.7907	(4.7554,4.8251)
δ_1	0.0067	(0.0061,0.0072)
γ	10.4560	(8.0198,12.8142)
β	-1.1373	(-1.5904,-0.6694)

Table 1: *Demand for cigarettes.* Summaries of the marginal posterior distributions.

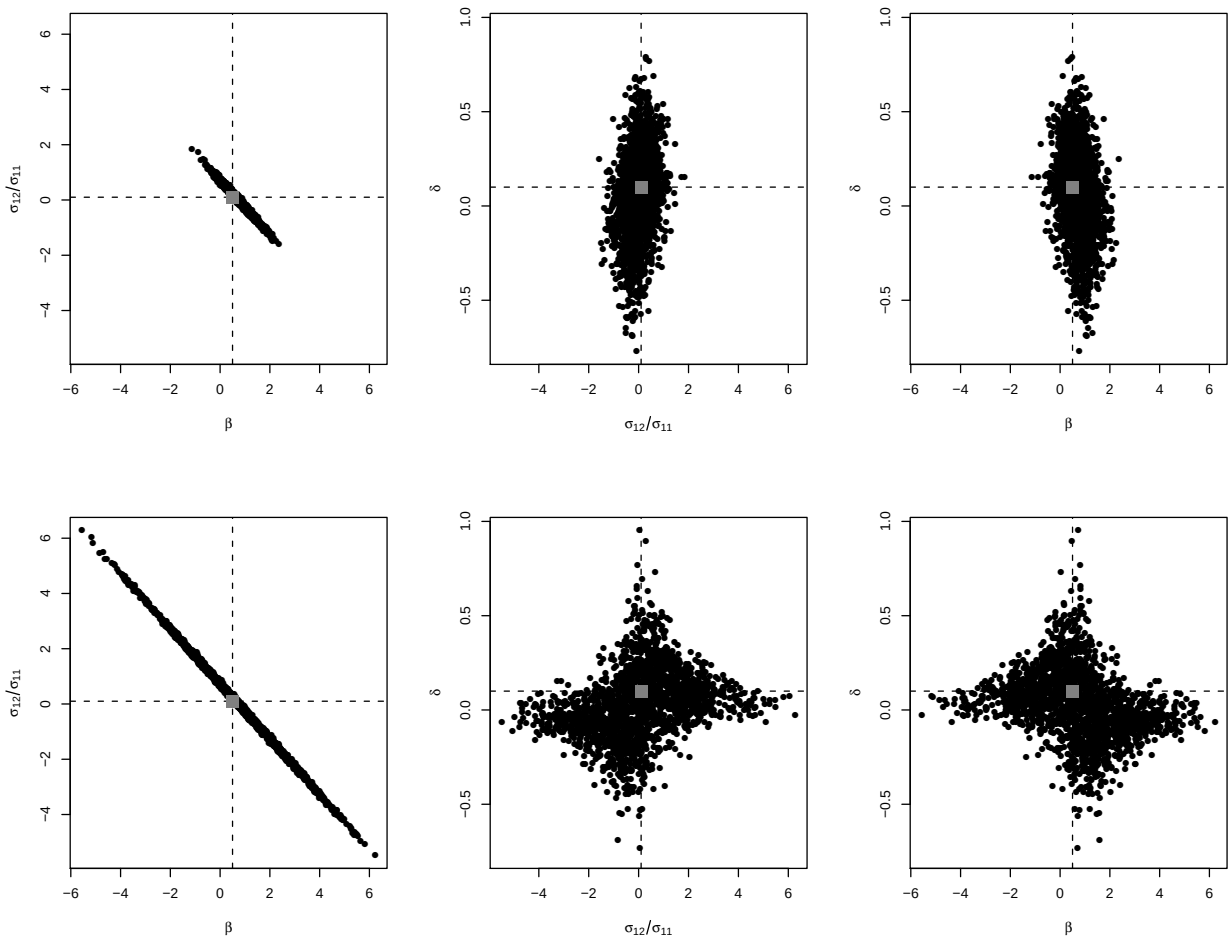

Figure 1: *Inverted Wishart prior.* $\Sigma \sim IW(v_0, \Sigma_0)$. Top row: $v_0 = 3$ and $\Sigma_0 = 3I_2$ (relatively vague prior). Bottom row: $v_0 = 0.00001$ and $\Sigma_0 = 0.00001I_2$ (non-informative prior). The independent priors on γ, β and the components of δ are all $N(0, 25)$. The Gibbs sampler is run for 1000000 draws (after discarding the initial 10000 draws) with every 1000th kept for posterior summaries.

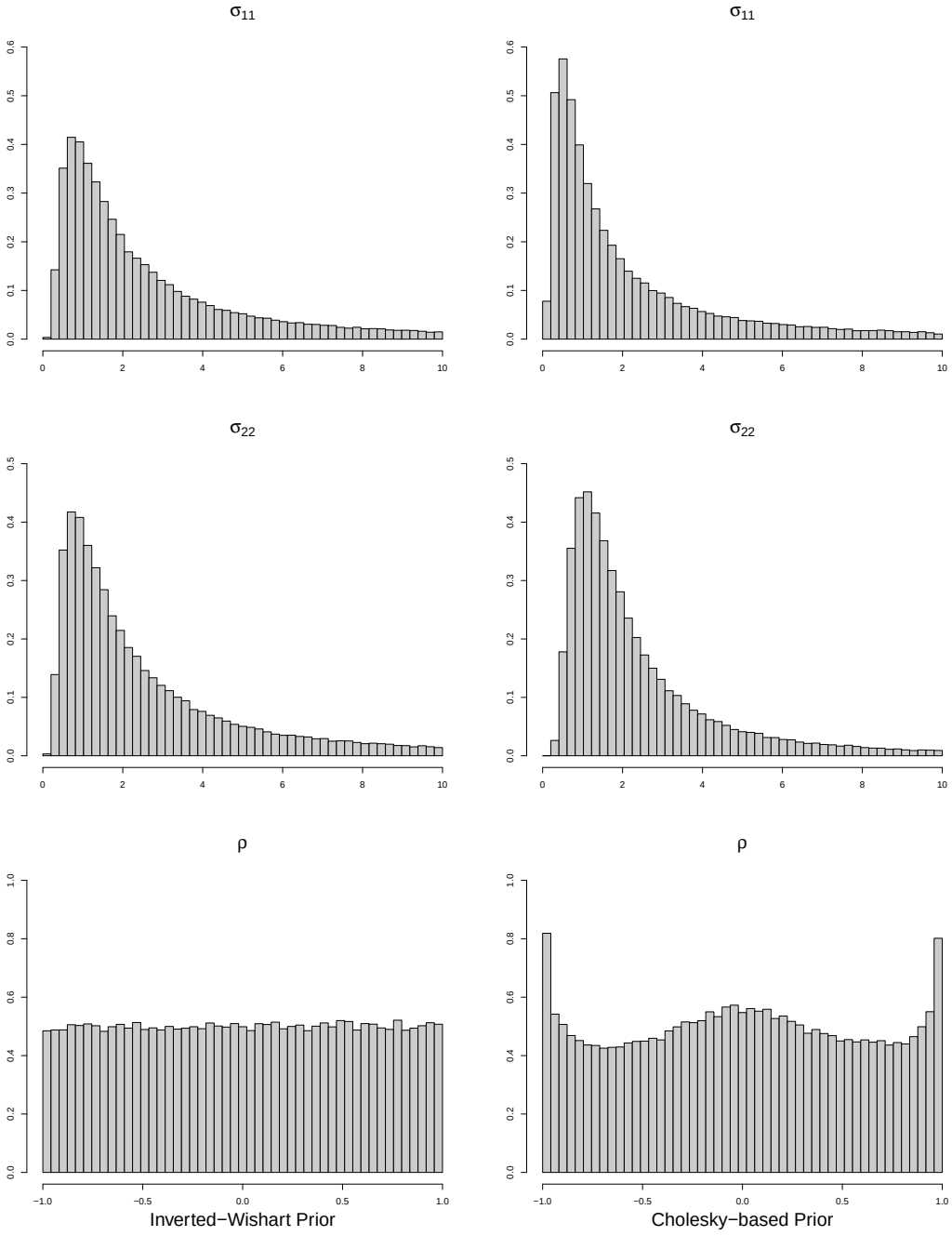

Figure 2: *Inverted Wishart and Cholesky-based priors.* $\Sigma \sim IW(3, 3I_2)$ (left column) and $\sigma_{11} \sim IG(0.75, 0.75)$, $\sigma_{2|1} \sim IG(3, 3)$ and $a_{21} \sim N(0, 0.7)$ (right column).

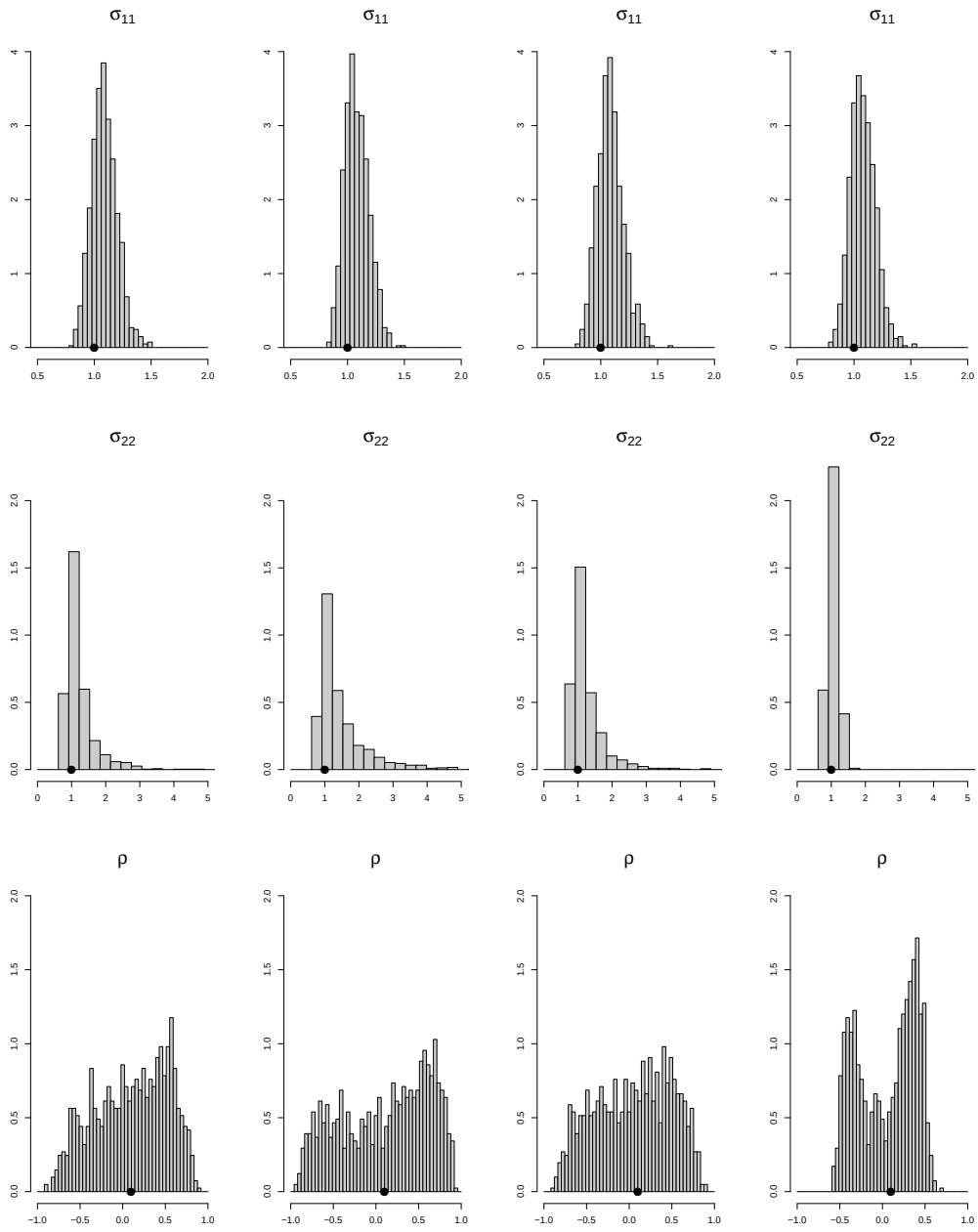

Figure 3: *Marginal posteriors.* Columns are histograms approximations to the marginal posterior distributions based on different prior specifications for (β, δ) and Σ. From left to right: $(\beta, \delta) \sim N(0, 25I_3)$ and $\Sigma \sim IW(3, 3I_2)$; $(\beta, \delta) \sim N(0, 25I_3)$ and Cholesky-based prior for Σ (see caption of Figure 2); $p(\beta, \delta) \propto ||\delta||(1+\beta^2)^{\frac{1}{2}}$ (see equation 10) and $\Sigma \sim IW(3, 3I_2)$; $p(\beta, \delta) \propto ||\delta||(1+\beta^2)^{\frac{1}{2}}$ and Cholesky-based prior for Σ.

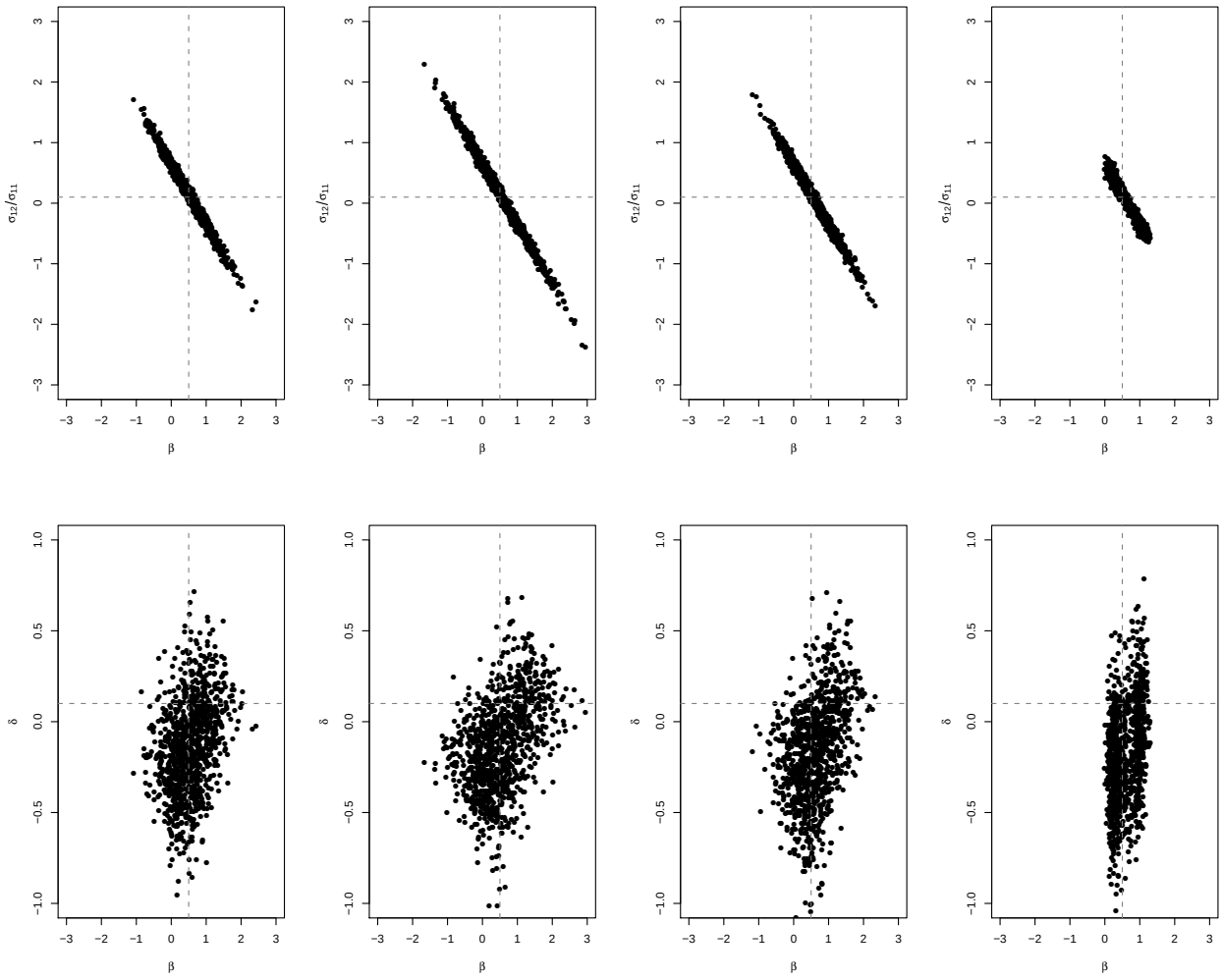

Figure 4: *Joint posterior of* $(\beta, \sigma_{12}/\sigma_{11})$ *and* (β, δ). Columns are as in Figure 3.

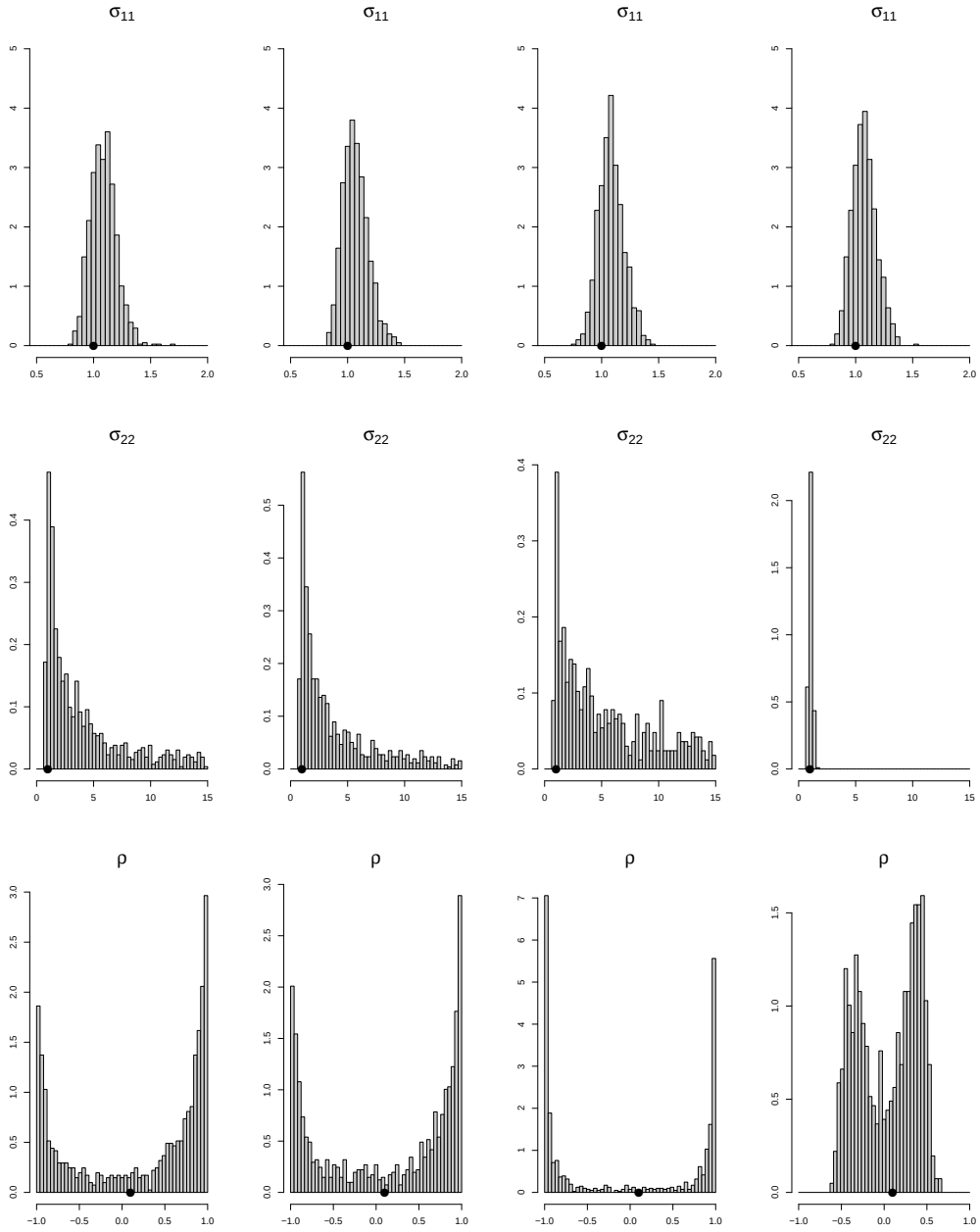

Figure 5: *Marginal posteriors.* Same as Figure 3, but based on more diffuse prior specification. For the inverted Wishart prior, $\Sigma \sim IW(0.00001, 0.00001 I_2)$, while for the Cholesky-based prior, $\sigma_{11} \sim IG(0.000001, 0.000001)$, $\sigma_{2|1} \sim IG(0.000001, 0.000001)$ and $a_{21} \sim N(0, 1000)$.

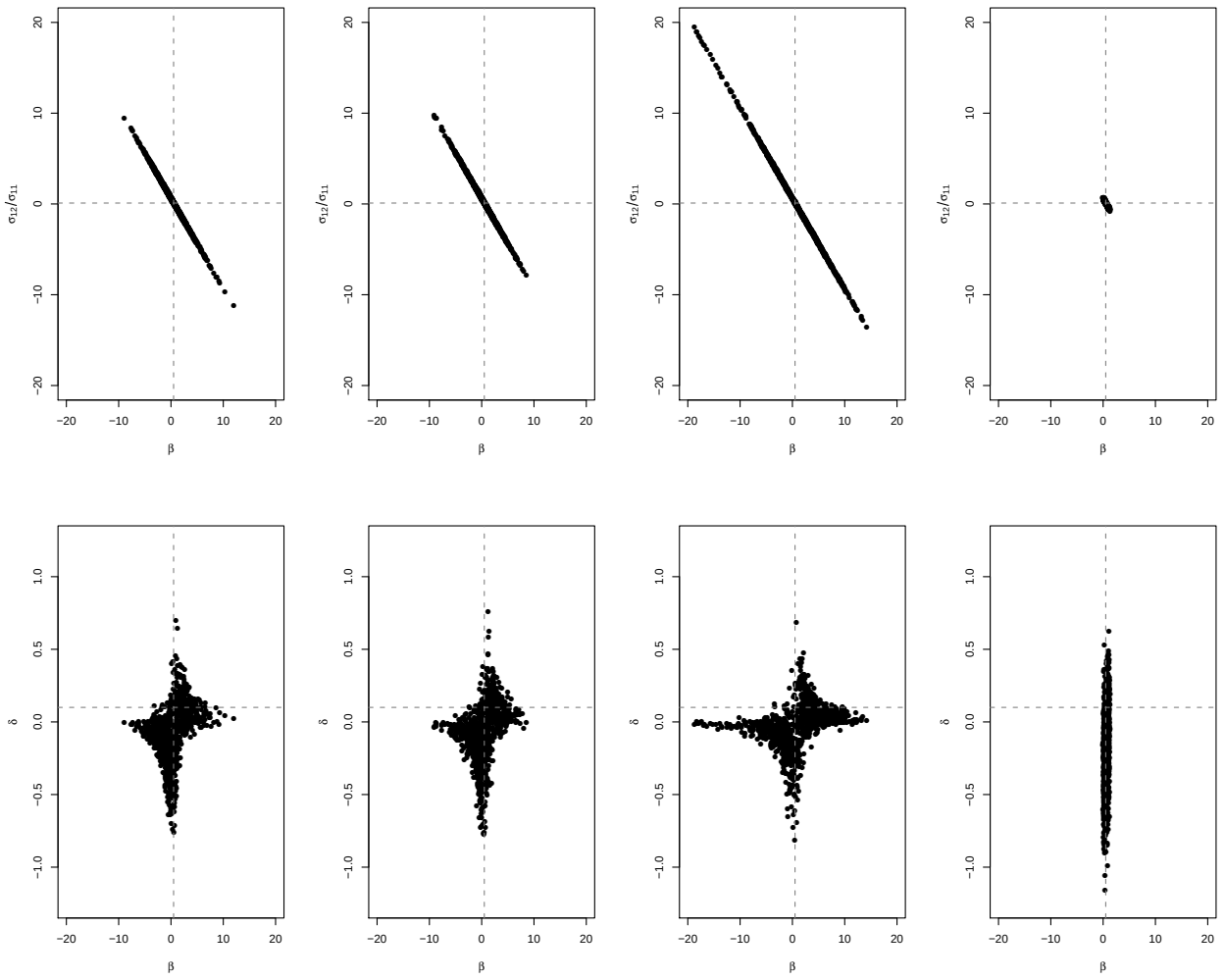

Figure 6: *Joint posterior of* $(\beta, \sigma_{12}/\sigma_{11})$ *and* (β, δ). Same as Figure 4, but based on the prior specification described in Figure 5.

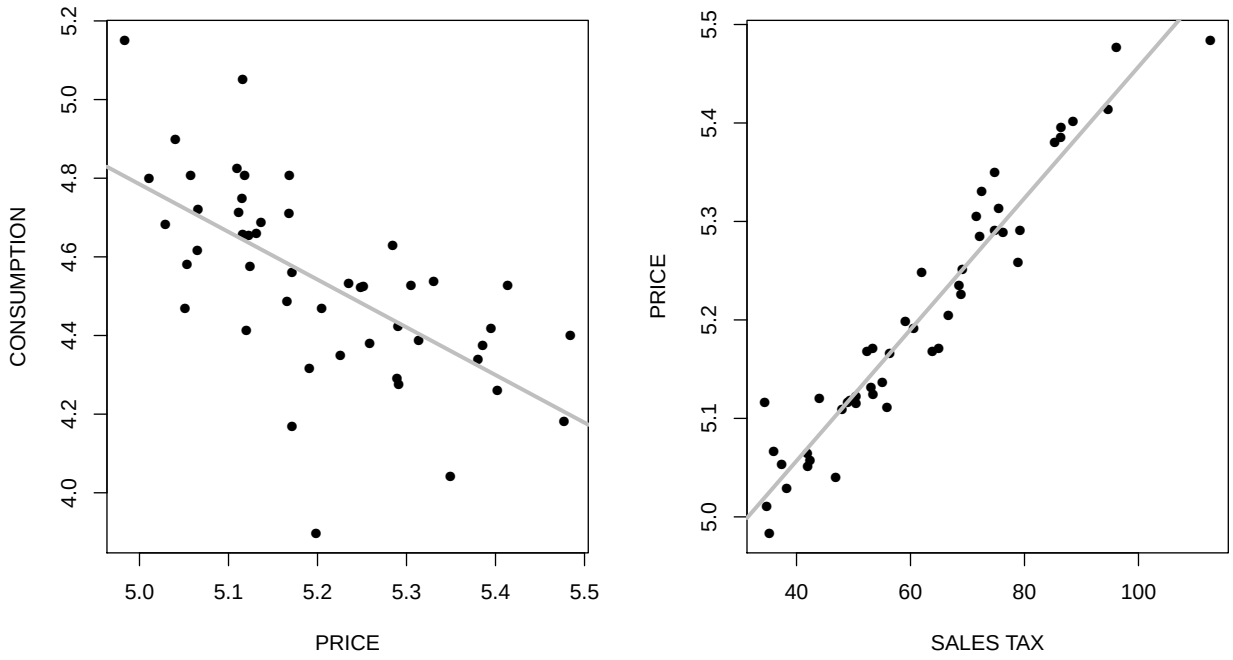

Figure 7: *Demand for cigarettes.* The data set consists of annual data for the 48 continental U.S. states for the year of 1995. `PRICE` is the logarithm of the average real price per pack of cigarettes including all taxes. `CONSUMPTION` is the logarithm of the number of packs of cigarettes sold per capita in the state. `SALES TAX` is the portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack (in real dollars, deflated by the Consumer Price Index).

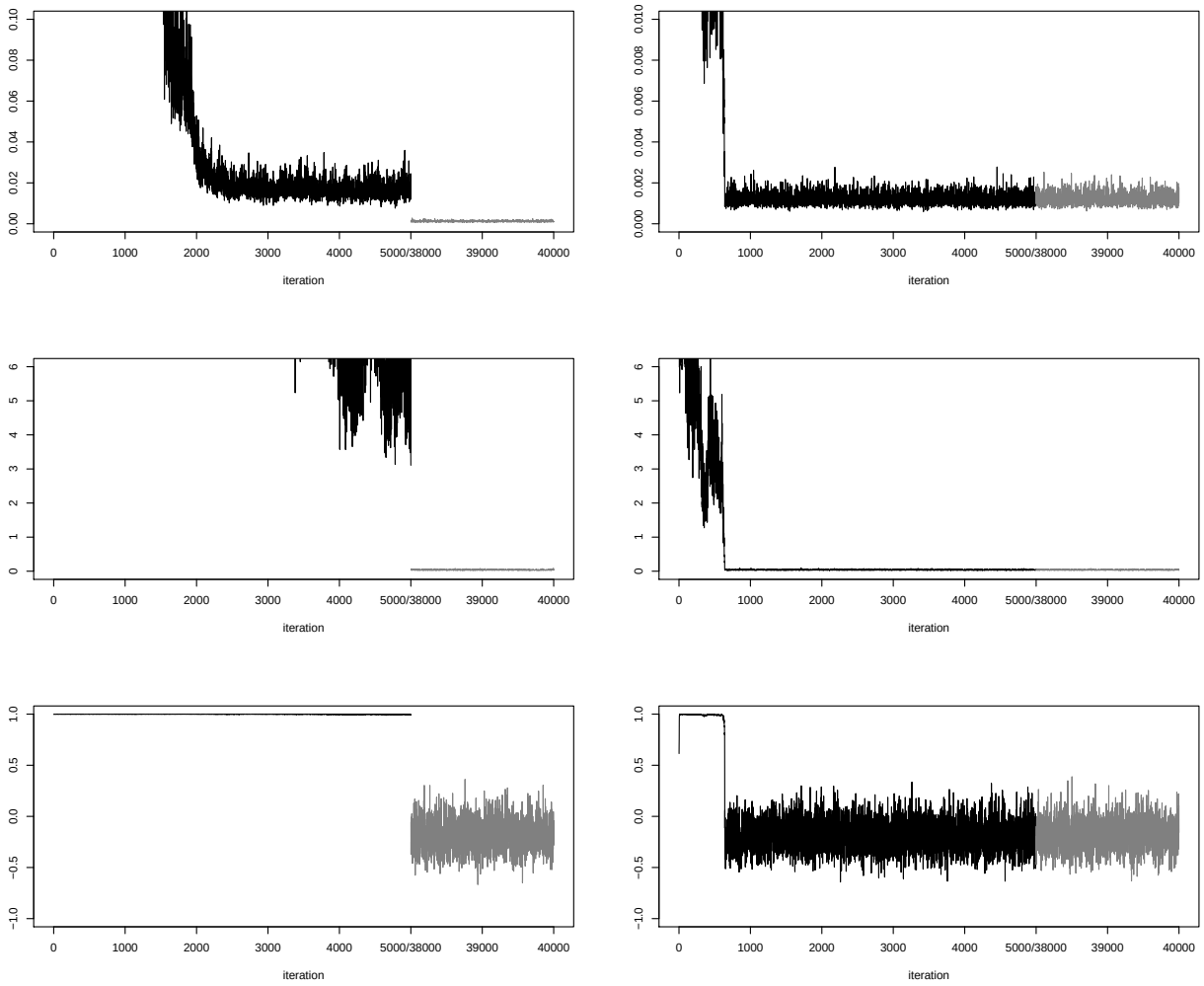

Figure 8: *Demand for cigarettes.* Trace plots of MCMC based on diffuse priors. Inverted Wishart prior (left panels) and Cholesky-based priors (right panels). After 5000 draws the MCMC based on the Inverted Wishart prior has not yet converged, while the MCMC based on the Cholesky-based prior converges after 1000 draws.

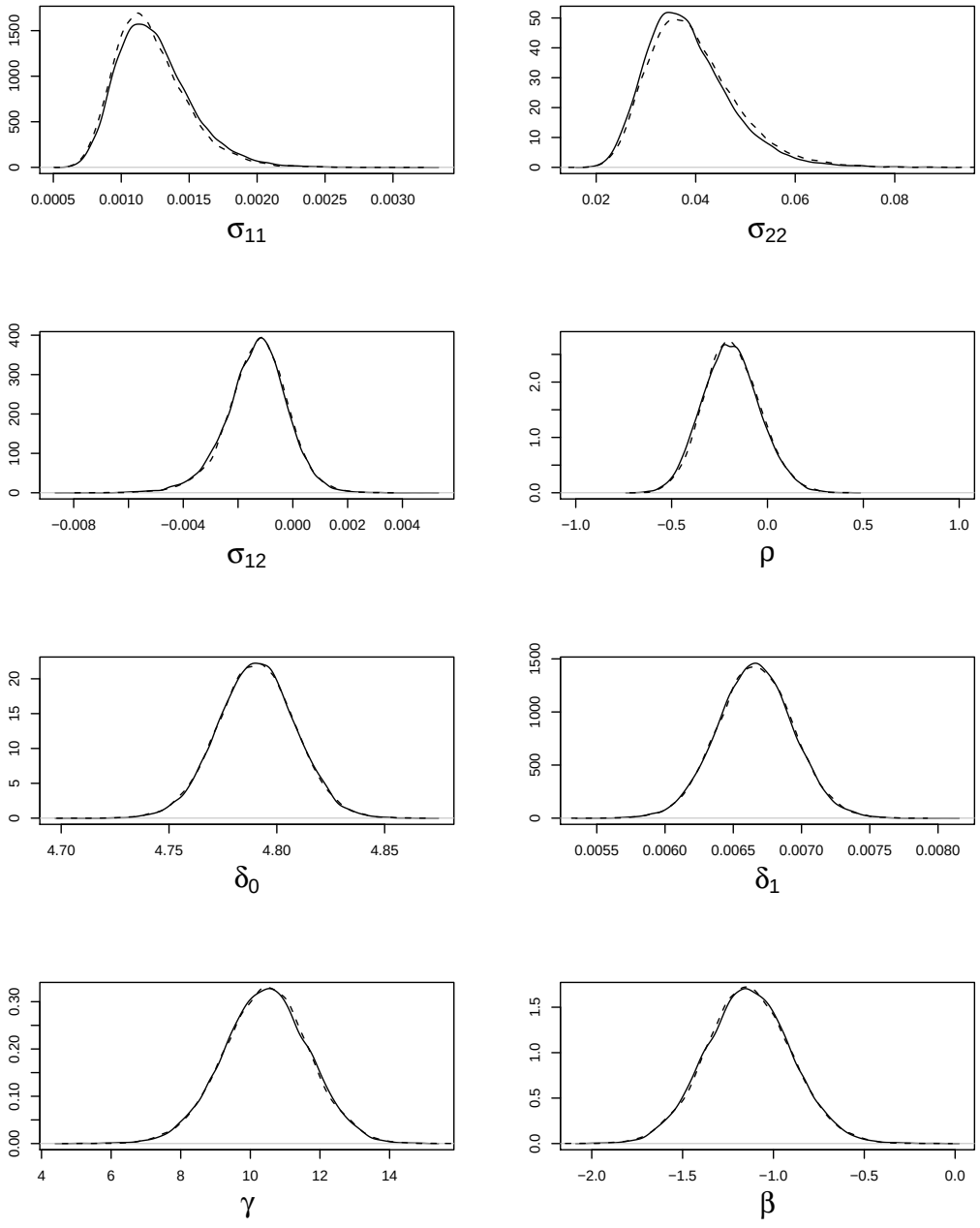

Figure 9: *Demand for cigarettes.* Marginal posterior densities based on diffuse priors. Inverted Wishart prior (solid lines) and Cholesky-based priors (dashed lines).

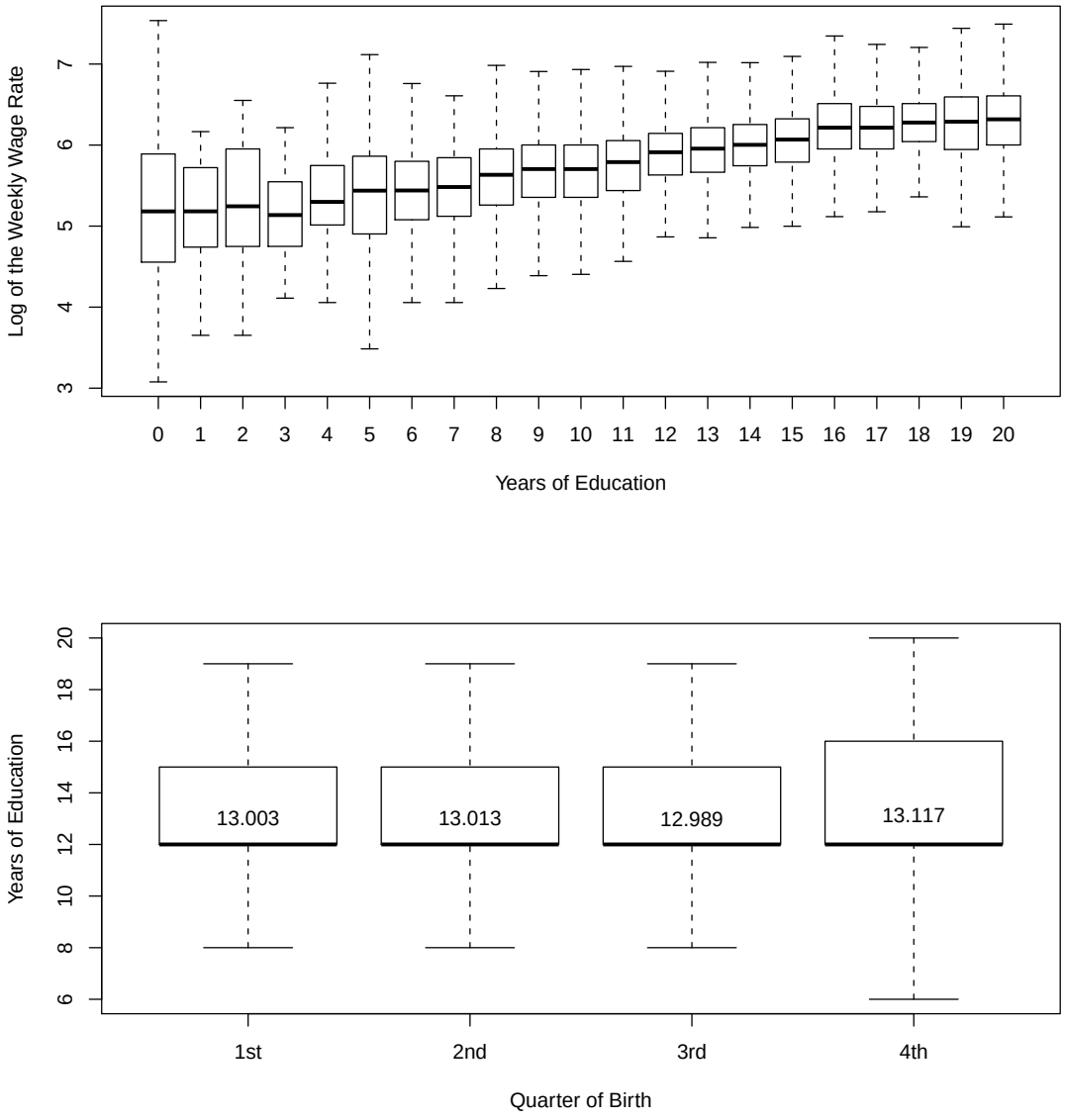

Figure 10: *Return to education.* Sample size of $n = 35,805$ men born in 1939. A subset of the data set analyzed by Angrist and Krueger (1991). Across quarters of birth, the median years of education is 12, corresponding to completion of high school. Means are slightly increasing from 1st to 4th quarter of birth, with the difference between 13.117 (4th quarter) and the other quarters ranges between 5.5 and 6.5 weeks of education.

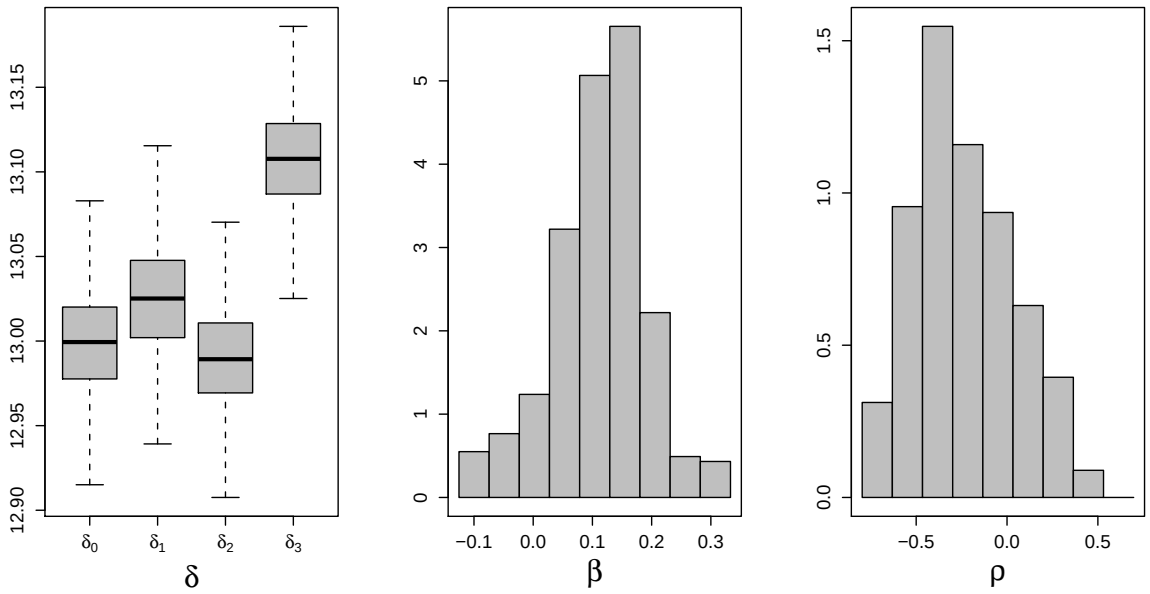

Figure 11: *Return to education.* Marginal posterior densities based on diffuse priors. Posterior means of δ_0, δ_1, δ_2 and δ_3 are 12.999, 13.025, 12.990 and 13.108, respectively, while $Pr(\beta < 0|\text{data}) = 7.7\%$ and $Pr(\rho < 0|\text{data}) = 73.4\%$.